Chapter 88
A Survey of Cloud-Based Services Leveraged by Big Data Applications

S. ZerAfshan Goher
National University of Sciences and Technology, Pakistan

Barkha Javed
National University of Sciences and Technology, Pakistan

Peter Bloodsworth
National University of Sciences and Technology, Pakistan

ABSTRACT

Due to the growing interest in harnessing the hidden significance of data, more and more enterprises are moving to data analytics. Data analytics require the analysis and management of large-scale data to find the hidden patterns among various data components to gain useful insight. The derived information is then used to predict the future trends that can be advantageous for a business to flourish such as customers' likes/dislikes, reasons behind customers' churn and more. In this paper, several techniques for the big data analysis have been investigated along with their advantages and disadvantages. The significance of cloud computing for big data storage has also been discussed. Finally, the techniques to make the robust and efficient usage of big data have also been discussed.

INTRODUCTION

The rapid advancement of the technology has proliferated the numbers of devices connected to the internet such as laptops, mobile devices, cameras, thermostats, wifi routers, electronic appliances, wireless sensors and more. According to a survey conducted by the Gartner, there are nearly fourteen billion devices connected to the internet so far and the number is expected to increase about twenty to hundred billion by the year 2020 (*Gartner Says 4.9 Billion Connected "Things" Will Be in Use in 2015*, 2014). The ever growing number of electronic devices using internet produces tremendous data streams and

DOI: 10.4018/978-1-5225-7501-6.ch088

aids to the concept of internet of things (IoT) where every day physical machines can be connected to the internet using wireless, RFID & sensor technologies (Gubbi, Buyya, Marusic, & Palaniswami, 2013). This extracted information from the raw data is exploited to draw smart decisions automatically such as the temperature and moisture sensor present in a wrist watch can help to adjust room temperature accordingly. Beside advantages, this growing interest in harnessing the internet has produced a massive amount of heterogeneous data that is difficult to handle by the conventional enterprise systems (Bandyopadhyay & Sen, 2011). This voluminous data is generated frequently at a high speed and is commonly referred to as big data. Big data is sheer volume of data that is so diverse and complex to be processed by a conventional database management system (DBMS) (Fan & Bifet, 2012). In order to handle this massive scale of data, data storage techniques need to be redefined. In this work, a survey of some of the important data storage techniques from the perspective of big data has been conducted.

The big data is originated from multiple sources and is therefore lacks a common standard pattern. Big data format can be categorized into three main classes which are: structured, semi-structured and unstructured data (Wu, Zhu, Wu, & Ding, 2014). The data in a structural approach is stored in the form of tables, where each table is composed of several rows and columns. The data stored in this format can be easily stored, analyzed and queried, whereas, the unstructured data lacks any fixed pattern and is found on a large scale such as website contents, blog entries, PDF files, radar data, satellites images, videos, social media data, business documents and more. The semi structured data lies in between; it is a type of structured data that does not own a rigid format. In semi-structured data, markers and tags are used to identify certain content of the data such as xml. Email is one of the examples of semi-structured data that has some fixed field (i.e. sender, receiver, date and time) along with some unstructured data (i.e. attachment and email body). The unstructured data is however generated more quickly than the other types of data and represents 80 percent of all the informatics data, whereas, structured and semi-structured schemes only captures five to ten percent of data (*Structured, semi-structured and unstructured data*, 2014). The management of this massive amount of data is beyond the capabilities of conventional DBMS and requires a scalable infrastructure that could mine data to gain useful insights. For instance, the text based sentiment analysis of the customer-centric data from social media helps to understand and to predict the customer behavior towards a particular product or campaign.

Data analytics is the core part of big data, sophisticated quantitative methods i.e. computational mathematics, data mining, machine learning, and artificial intelligence are used to perform analysis and to discover hidden patterns for enhancing business productivity (Fisher, DeLine, Czerwinski, & Drucker, 2012). Big data requires a scalable architecture and high processing power to operate large volume of distributed data. However, the high adoption cost associated with the hardware procurement makes it difficult for enterprises to leverage the up-to-date tools and technologies to process volumetric data. Cloud computing is a viable option in such scenarios as it reduced the resource acquisition cost by implementing pay-as-you-go model (Vaquero, Merino, Caceres, & Lindner, 2009). The elasticity of cloud technology makes it a good candidate to manage the big data applications. Cloud facilitates on the fly expansion of underlying infrastructure i.e. when the resource demands increases beyond the capacity of an enterprise then the cloud resources can be reserved with minimum effort. This relief user from the burden of over-provisioning where an enterprise invests heavily to build up its infrastructure, however, in case of low resource demands the added capacity goes underutilized which reduced revenue streams. Cloud-oriented big data analytics provides pre-built setup to support data-driven decision making (Buyya, Yeo, Venugopal, Broberg, & Brandic, 2009).